Exploring the Determinants of Medical Insurance Costs

Drew Levin dslevin2@wisc.edu
Sakshi Shah sashah7@wisc.edu
Rachel DeGeorge rdegeorge@wisc.edu
Ian Jacobson idjacobson@wisc.edu

1. Introduction

Medical Insurance is often an expensive yet valuable investment to make. The confusion around costs and premiums is difficult to navigate. In order to better understand what changes the costs of insurance, this report revolves around a dataset that includes many demographic and lifestyle variables including age, sex, body mass index (BMI), number of children, smoking status, geographical region, and the corresponding insurance charges, as illustrated in the sample data figure below.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure 1. Sample data.

Through our analysis, our study aims to answer: What factors contribute to higher medical insurance costs? How do these costs vary geographically?

In our analysis, we used machine learning models such as: Linear Regression, Decision Tree Regressor, and Gradient Boosting. Our findings showed a high accuracy score for the Gradient Boosting model with an emphasis on the smoker, age, and BMI features as the most noteworthy factors on medical costs.

2. Data Visualization

2.1. Dataset

The dataset contains 2,772 entries with 7 columns, culminating in individual medical costs billed by health insurance. The variable we are focusing on, "charges," shows the insurance cost associated with each individual. The dataset includes a wide age range, though it features a significant number of individuals between 18 to 22.5 years old. Of those individuals, most are male. Most have less than three children, and the bulk falls within a BMI range indicating overweight to moderate obesity (29.26 to 31.16). The data encompasses four regions (northeast, northwest, southeast, and southwest).

In order for our models to better represent the data, we converted the binary features (sex, smoker) into numerical data to standardize the dataset. Taking the analysis further, we also split up the dataset by region in order to determine which places in the United States pay the most for health insurance.

2.2 Feature Importance

To gain a little insight into the potential features that would be important for predicting charges, we computed permutation feature importance on the training set. The figure below highlights smoker, age, and BMI as being very influential.

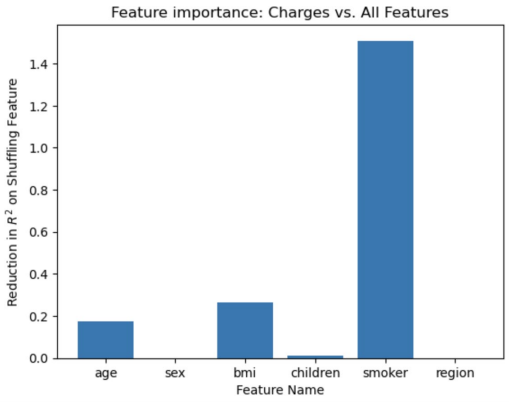

Figure 2. Feature Importance.

2.3. Linear Regression

We started our analysis with a simple Linear Regression model. An interesting observation from the analysis is the impact of smoking on insurance costs across different regions. Smoking status influences insurance charges far more than any other variable, with coefficients varying by region. This variability can be explained by regional differences in the pricing policies of insurance companies or possibly different health risk profiles that are associated with smokers in these regions.

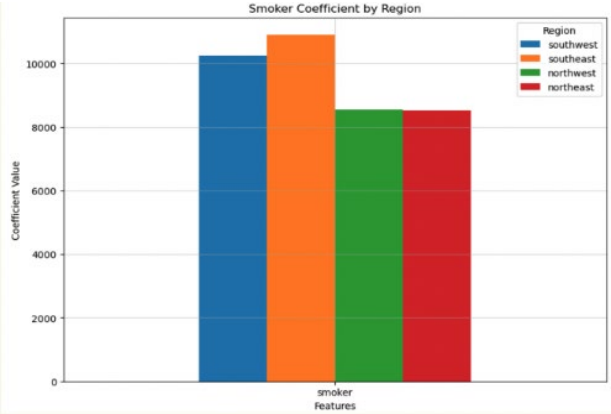

Figure 3. Simple Linear Regression of Smoker Charges by Region.

Another notable observation was the Sex coefficients of the Linear Regression. It can be observed that (1: Female, 0: Male) women on average pay more for insurance compared to men, except in the Northeast.

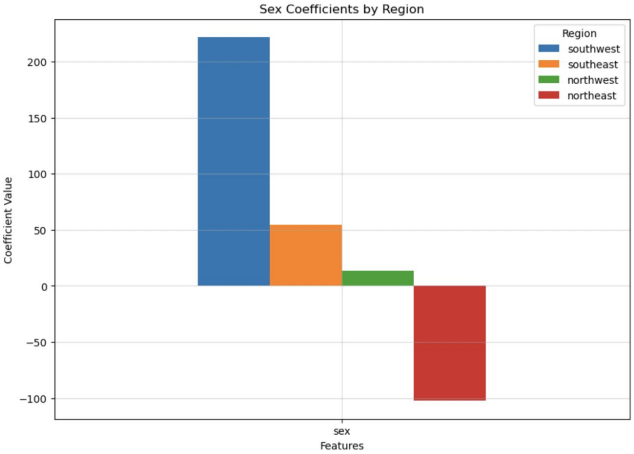

Figure 4. Simple Linear Regression of Sex Charges by Region.

3. Experiments

3.1. Model Implementation: Decision Trees and Gradient Boosting

We used a Decision Tree Regressor to evaluate the non-linear interactions between the predictors. The model's effectiveness with insurance rate prediction was shown through a plot comparing actual charges against predicted charges.

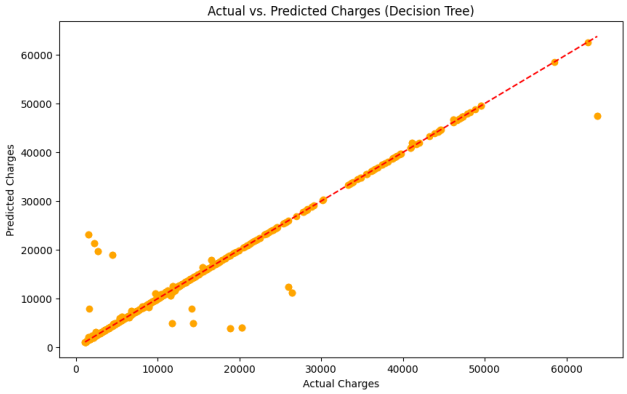

Figure 5: Actual vs. Predicted Charges using the Decision Tree model

We also used Gradient Boosting. The actual vs. predicted charges plot for this model showed even tighter alignment, indicating better performance.

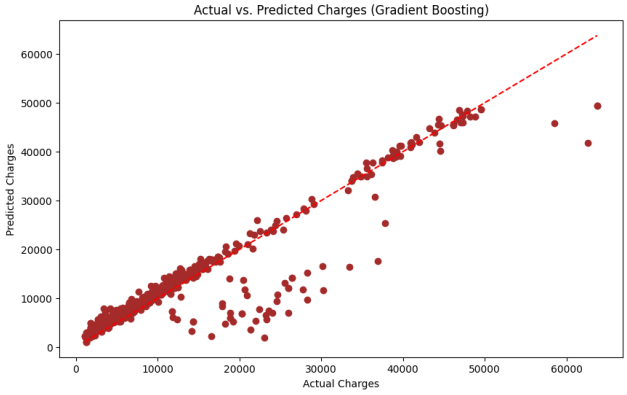

Figure 6: Actual vs. Predicted Charges using Gradient Boosting

3.2. Model Validation

We validated these models using a split of 80% training data and 20% testing data.

3.3. Performance Metrics

Model performance was evaluated using R^2 and RMSE. Both metrics confirmed that our models were very accurate, especially the Gradient Boosting model with a higher R^2 value and lower RMSE compared to the others.

- **Linear Regression:** R^2=0.746
- **Decision Tree Regressor (max_depth = 3)**: R^2=0.856
- **Gradient Boosting:** R^2 = 0.884

4. Conclusion

In this project we used three different machine learning models to determine the medical insurance costs. The gradient boosting method was the most successful, providing the most accurate predictions and giving insight into which predictors were the most influential.

Although the analysis was insightful, it would be interesting to study how different factors such as economic status or existing medical conditions would affect insurance rates.

5. Contributions

Member	Proposal	Coding	Presentation	Report
Drew	1	1	1	1
Sakshi	1	1	1	1
Rachel	1	1	1	1
Ian	1	1	1	1

References

Vyas, Rahul. "Medical Insurance Cost Prediction." *Kaggle*, 2022, www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction/data. Accessed 14 April 2024.